Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

DIPLOMOVÁ PRÁCE

Automatizovaný výběr dat pro trénování modelu porozumění mluvené řeči

Automatized data selection for training the spoken language understanding model

Plzeň, 2014

Autor: **Bc. Markéta Jedličková**

Vedoucí práce: **Ing. Jan Švec, Ph.D.**

Prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne _____ _____

podpis

Poděkování

Ráda bych poděkovala Ing. Janu Švecovi, Ph.D. za cenné rady, věcné připomínky a vstřícnost při konzultacích a vypracování diplomové práce.

Anotace

Tato diplomová práce představuje návrh několika strategií pro automatizovaný výběr trénovacích dat modelu porozumění mluvené řeči. Nejprve jsou obecně popsány hlasové dialogové systémy a jejich souvislost s oblastí výběru dat pro trénování modelu. Dále je popsán hierarchický diskriminativní model používaný v této práci k ověření navržených metod. Následuje přehled nově vyvinutých strategií zabývajících se náhodným výběrem dat a výběrem dle hodnoty posteriorní pravděpodobnosti, F-skóre a míry V_u. Pro tyto metody je provedeno experimentální ověření a jsou po natrénování modelu vyhodnoceny za pomoci konceptové přesnosti. Závěrečná kapitola shrnuje přínos navržených strategií pro výběr trénovacích dat modelu porozumění mluvené řeči a jejich využití v praxi.

Klíčová slova: hlasové dialogové systémy; porozumění mluvené řeči; výběr dat pro trénování modelu

Annotation

This master's thesis presents different methods for automated selection of training data for spoken language understanding. First, the spoken dialog systems are described in relation to area of training data selection. Then the hierarchical discriminative model is described. This model is used to verify the proposed methods. The following is an overview of newly developed strategies dealing with random data selection and data selection according to the value of the posterior probability, F-score and the rate of V_u. For these methods is performed experimental verification and their impact on the concepts accuracy. The final chapter summarizes the benefits of proposed strategies for the training data selection and its use in practice.

Keywords: spoken dialog system; spoken language understanding; selection of training data

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Akademický rok: 2014/2015

Studijní program: Aplikované vědy a informatika
Forma: Prezenční
Obor/komb.: Kybernetika a řídicí technika (KŘT)

Podklad pro zadání DIPLOMOVÉ práce studenta

PŘEDKLÁDÁ:	ADRESA	OSOBNÍ ČÍSLO
Bc. JEDLIČKOVÁ Markéta	Pod Školou 10, Plzeň - Červený Hrádek	A12N0163P

TÉMA ČESKY:

Automatizovaný výběr dat pro trénování modelu porozumění mluvené řeči

NÁZEV ANGLICKY:

Automatized data selection for training the spoken language understanding model

VEDOUCÍ PRÁCE:

Ing. Jan Švec - NTIS

ZÁSADY PRO VYPRACOVÁNÍ:

1. Nastudujte si problematiku automatického rozpoznávání a porozumění řeči.
2. Navrhněte postup pro přetrénování modelu porozumění. Zaměřte se na různé možnosti automatizovaného výběru nových trénovacích dat.
3. Jednotlivé možnosti výběru trénovacích dat experimentálně ověřte.

SEZNAM DOPORUČENÉ LITERATURY:

Literaturu dodá vedoucí diplomové práce.

Podpis studenta: ..... **Datum:** .....

Podpis vedoucího práce: ..... **Datum:** .....

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

Zkratka	Popis
ASR	Automatic Speech Recognition, automatické rozpoznání řeči
cAcc	Konceptová přesnost (concepts accuracy) (str. 16)
HDM	Hierarchický diskriminativní model
STC	Semantic Tuple Classifiers, klasifikátory sémantických n-tic
SVM	Support Vector Machine
TIA	Telefonní Inteligentní Asistentka (řečový korpus)
WFST	Weighted Finite State Transducers, vážený konečný stavový transducer

Kapitola 1

Úvod

Tato diplomová práce je zaměřena na automatizovaný výběr dat pro trénování modelu porozumění mluvené řeči. Oblast dialogových systému zaznamenala v posledních letech nebývalý rozvoj. Ke zlepšení dochází i v jejich využití v praxi. Například v rámci výzkumného projektu MPO TIP FR-TI1/518, vznikla telefonní inteligentní asistentka (TIA). Tento systém by měl být schopný komunikovat v oblasti administrativy běžnou řečí a vykonávat základní požadavky spojených s touto oblastí. Tento projekt byl řešený firmou SpeechTech s.r.o ve spolupráci s katedrou kybernetiky Západočeské univerzity v Plzni. Tato práce čerpá z korpusu[1] tohoto projektu.

Dialogový systém lze popsat jako počítačový model komunikující s lidmi za pomoci porozumění mluvenému jazyku. První pokusy o komunikaci probíhající mezi člověkem a strojem ve formě dialogu začaly vznikat prostřednictvím textu. Tento způsob komunikace se objevil v šedesátých letech minulého století [2]. Jedním z prvních průkopníků byl Alan Turing, který v práci z roku 1950, nazvané *„Computing Machinery and Intelligence“* [3], formuloval myšlenku později pojmenovanou jako *„Turingův test“*. Tento test měl za úkol prověřit, zda je možné vytvořený dialogový systém rozpoznat od člověka, či nikoliv.

Navržené dialogové systémy se snaží o imitaci určitého typu chování člověka. Pokud stroj nedokáže odpovědět na otázku kladenou uživatelem, odpoví na ní další otázkou. Tento postup umožňuje pokračovat v komunikaci i s omezeným zdrojem znalostí [4]. Rozvoj těchto systémů je komplexní záležitost, která je výzvou k výzkumu technologií zabývající se komunikací člověka s počítačem.

[1] *Řečový korpus*, soubor řečových záznamů jehož nedílnou součástí je tzv. anotace, tj. symbolická reprezentace řeči [1, s. 459]

V současné době se stává výskyt hlasových dialogových systémů v praxi stále častější. Komunikace probíhající prostřednictvím hlasu začíná být běžná jak v počítačích tak v mobilních telefonech. Zatím se ovšem nejedená o zcela plnohodnotný dialog, ale vývoj těchto technologií jde rychle kupředu a je předpoklad, že v následujících letech budou dialogové systémy stále rozšířenější. Postupně s jejich rozšířením se tyto systémy stanou uživatelsky přirozenějšími. K tématu současného stavu dialogových systémů se hodí následující citát:

> *We can only see a short distance ahead, but we can see plenty there that needs to be done.*
>
> *Alan Turing.* [3]

Cílem této práce je posoudit jakým způsobem lze ovlivnit výslednou konceptovou přesnost vhodným výběrem dat složených z uskutečněných dialogů. Práce popisuje čtyři různé metody a výsledky jejich použití pro tento model. Při návrhu strategií výběru dat pro trénování modelu porozumění mluvené řeči byla využita disertační práce [5], popisující hierarchický diskriminativní model, na kterém jsou ověřeny navržené strategie.

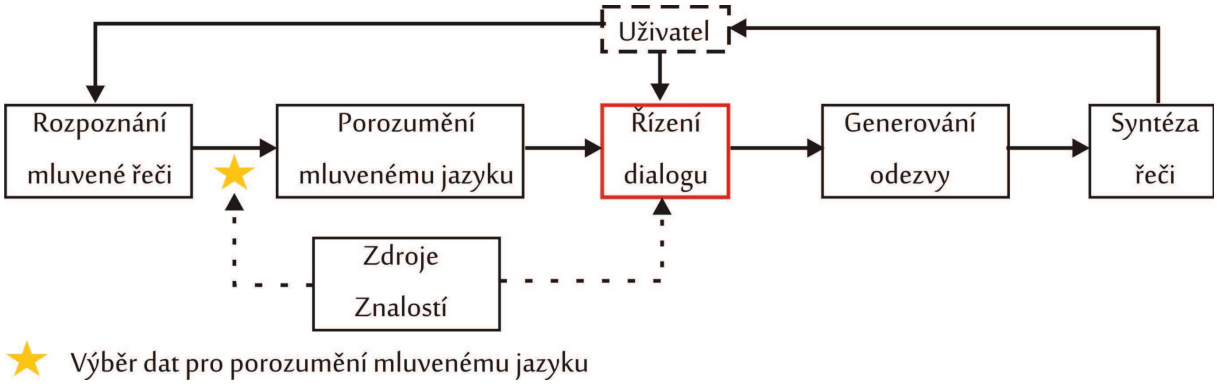
Tato práce je rozdělena do šesti kapitol. První kapitolou je: *„Úvod“*, kde je uveden cíl, téma a struktura práce. V druhé kapitole s názvem: *„Hlasové dialogové systémy“*, jsou teoreticky představeny dialogové systémy a jednotlivé moduly, ze kterých se každý takový systém skládá. Kapitola: *„Hierarchický diskriminativní model“*, popisuje nově vytvořený model [5], který se používá k ověření jednotlivých metod popsaných v kapitole: *„Strategie výběru dat“*. V páté kapitole: *„Experimentální ověření“*, jsou jednotlivé metody ověřeny a jsou v ní uvedeny jejich výsledky. Šestá kapitola: *„Závěr“*, uvádí shrnutí diplomové práce a dokládá přehled dosažených výsledků.

Kapitola 2

Hlasové dialogové systémy

Hlasové dialogové systémy slouží ke komunikace mezi člověkem a počítačem prostřednictvím hlasu. Pod touto definicí se skrývá několik možností realizace dialogových hlasových systémů, od jednoduchých, které se skládají pouze z několika příkazů, až ke komplexním, umožnující plynulou konverzaci. Při návrhu systému je nutné splnit několik požadavků, zajišťující správný chod systému. Patří sem zejména *vysoká robustnost*, *dobrá srozumitelnost* a *rychlá odezva* [1].

K zajištění úspěšného průběhu komunikace je nutné vytvořit vhodný návrh systému, který je tvořen z několika modulů obstarávající následující funkce: *řízení dialogu*, *generování odezvy*, *porozumění mluvenému jazyku*, *rozpoznání a syntéza řeči*. Další důležitý požadavek je zajištění modelování potřebných **zdrojů znalostí**, které simulují apriorní znalosti o světě. Zjednodušené blokové schéma typického hlasového dialogového systému je možné vidět na obr. 2.1.



Obrázek 2.1: Model hlasového dialogocého systému.

V modelu na obr. 2.1 je pomocí žluté hvězdy vyznačena část, ve které jsou aplikovány jednotlivé metody zvolené k výběru dat pro porozumění mluvené řeči. Tato práce se zabývá především touto specifickou oblastí. K ověření navržených metod byl použit **hierarchický diskriminativní model**[5], blíže specifikovaný v kapitola 3.

O vytvoření systému, který by byl schopný rozpoznat libovolného řečníka mluvícího jím zvoleným jazykem, se vědci snaží již přes padesát let. Řešení se zatím nepodařilo nalézt, protože je nejprve potřeba překonat několik zásadních problémů. Jedná se o odlišnosti hlasů jednotlivých řečníků, akustické prostředí, odlišnost hlasu řečníka v jednotlivých situací či jev *koartikulace*[2]. Tato problematika spadá do modulu, který se nazývá **rozpoznání mluvené řeči**.

Pro hlasové dialogové systémy je rozhodujícím blokem **porozumění mluvenému jazyku**. Pomocí jedné z navržených strategií, kterými se tato práce zabývá, jsou z množiny uskutečněných dialogů vybrána data, která jsou ohodnocena jako nejvíce vhodná k anotaci. Jejich vhodnost k anotaci je dána hodnotou výběrového faktoru jednotlivých strategii, více v kapitole: *„Strategie výběru dat“*. Po získání dat se tento modul snaží o přiřazení kontextu získané posloupnosti slov se zohledněním dalších známých informací. Význam vstupní promluvy je získáván za pomoci tří složek: **syntaktické analýzy**, **reprezentace znalostí** a **interpretace významu**.

Pokud by výsledek rozpoznání a porozumění mluvenému jazyku nebyl dostatečně věrohodný, ovlivnilo by to výstup systému, který by byl zatížen mnoha nedorozuměními a chybami. V posledním desetiletí byl často studován **pravděpodobnostní model** rozpoznání vstupních dat, který byl shledán vyhovujícím. Vylepšený pravděpodobnostní model je také použit v hierarchickém diskriminitavním modelu.

Rozhodovací proces nazvaný jako **řízení dialogu** probíhá v prostoru stavů, akcí a strategií dialogu. Veškerá znalost systému získaná vzájemnou interakcí mezi uživatelem a systémem, včetně reprezentace stavů a úlohy řešené dialogem se nazývá **stav dialogu**. Tento stav se mění při uskutečnění **akce dialogu**. Mezi jednotlivé akce dialogu patří jakýkoliv akční zásah, jako je například obdržení informace od uživatele. **Strategie dialogu** specifikuje budoucí akci, která je vybrána pro každý stav, kterého je v dialogu

[2] *Koartikulace*, tento jev spočívá ve změně fonetické vlastnosti začátku a konce slova v závislosti na kontextu okolních slov [1, p. 195]

dosaženo. Strategie by měla být navržena tak, aby uspokojila požadavky uživatele. Za získání požadovaných informací, vybrání vhodné strategie a odpovídajících akcí je odpovědný **dialogový manažer**, který organizuje chod systému tak, aby bylo dosaženo cílového chování.

Generování odezvy představuje proces zajišťující sdělování potřebné informace uživateli dialogového systému prostřednictvím hlasového výstupu. Jde především o zvolení vhodné struktury a formy odpovědi obsahující výstupní informaci se snahou o co nejpřirozenější komunikaci mezi uživatelem a systémem. Složitější systémy využívají pro generování odezvy také znalost **míry důvěry**, která může být získávána v průběhu promluvy s uživatelem a zajišťuje zvýšení přirozené komunikace.

Závěrečným procesem je **syntéza řeči** zajišťující umělé vytvoření řeči pomocí počítače. Tento proces si klade za cíl přirozenější komunikaci mezi člověkem a počítačem. Zařízení, zajišťující tento proces, se nazývá **syntetizér řeči**.

Využití dialogových systémů je používáno především v případech, kdy uživatel není schopen ovládat technologii jiným způsobem, nebo je zaměstnán jinou činností. Mezi tyto činnosti patří například řízení auta. Další využití lze nalézt v technologických či lékařských aplikací. V poslední době jsou rozvíjeny možnosti kombinace různých metod pro ovládání těchto zařízení. Jednou z oblastí jsou i multimodální dialogové systémy, které kombinují vizuální a řečovou interakci. Pokud jsou jednotlivé funkce modelu správně navrženy a jsou zohledněny systémové požadavky, splníme tak nutnou podmínku pro vytvoření uživatelsky přívětivého hlasového dialogu. I po získání dobře navrženého dialogového systému je stále velmi důležitý přístup uživatele.

V současné době se lidé komunikace se strojem stále trochu obávají. Často mají sklon k tomu, aby systém podceňovali, nebo naopak, jsou jejich nároky příliš vysoké. Komunikace se strojem jim poté může připadat nepřirozená. Každým dnem se tyto problémy zmenšují, také například díky nadčasovým filmům, které jsou příčinou toho, že si lidé na myšlenku plnohodnotného rozhovoru s počítačem pomalu zvykají a připadá jim čím dál tím více reálnější. Se vznikem nových technologií přicházejí stále nové příležitosti na zdokonalení již existujících metod a díky tomu se možnosti hlasových dialogových systémů posouvají stále dál.

Kapitola 3

Hierarchický diskriminativní model

Tato kapitola popisuje hierarchický diskriminativní model [5], který je v této práci použit k ověření navržených strategií výběru trénovacích dat. Tento model je schopen zpracovat vstup v podobě slovní nebo fonémové mřížky či řetězce a poté vyhodnotit více významových hypotéz. Model je určen pro použití v hlasových dialogových systémech a je rozšířením modelu STC[3].

Mezi výhody HDM patří například možnost získání více vstupních hypotéz, generování výstupního sémantického stromu založeného na heuristice či trénování z neurčitého vstupu. Pro popis modelu je použita terminologie vycházející z perceptronových neuronových sítí. Model je rozčleněn do tří základních vrstev:

- **Vstupní vrstva**, slouží k vyčíslení hodnot racionálních jádrových funkcí mezi jednotlivými promluvami z trénovací množiny použitých pro trénování a predikci.
- **Skrytá vrstva**, převážně odpovídá modelu STC. Jejím cílem je vytvoření vektoru vzdáleností vstupní promluvy k oddělovacím nadrovinám binárních SVM.
- **Výstupní vrstva**, předpovídá pravděpodobnosti jednotlivých sémantických pravidel, ze kterých je sestaven výstupní abstraktní sémantický strom.

Model se může skládat ze dvou nebo tří vrstev propojených dopřednou vazbou. Zjednodušené schéma modelu na obr. 3.1 [5] popisuje jednotlivé vrstvy a jejich parametry.

[3] *Semantic Tuple Classifiers*, klasifikátory sémantických n-tic [6].

Při popisu hierarchického diskriminativního modelu jsou použita následující označení:

- *Vážený konečný transducer* - jednotící prvek spojující symbolické a statistické paradigma [5, s. 2], neboli jeden ze způsobů indexace slovních mřížek.
- *Vážený konečný akceptor* - speciální případ váženého konečného transduceru. Nedefinuje relaci mezi vstupními a výstupními řetězci, ale pouze vahou oceňuje cestu akceptorem [5, s. 34].
- *Vážené konečné automaty* - neboli vážené konečné *akceptory* nebo *transducery*, v práci HDM použité pro reprezentaci vstupních promluv.
- *Slovní mřížka* - data ve formě acyklického váženého akceptoru získané na výstupu systému automatického rozpoznávání řeči.
- *Fonémová mřížka* - data ve formě acyklického váženého akceptoru získané z fonémového rozpoznávače.
- *Řetězec* - automat ve kterém existuje právě jedna cesta nenulové délky [5, s. 33].
- *Sémantická entita* - odpovídá konkrétní slovní realizaci v dané promluvě. Typem sémantické entity může být například datum, čas, jméno apod.[5, s. 11]
- *Koncept* - značka (tag), odlišující různé významy promluv a různé třídy entit v rámci jednotlivých promluv [5, p. 10]. Na základě korpusu TIA byli v rámci výzkumného projektu například definovány koncepty jako SCHUZKY, DATUM, JMENO apod.
- *Sémantická n-tice* - podmnožina cesty z kořene sémantického stromu k jakémukoliv dalšímu uzlu. Například v anotaci VYTVOR(SCHUZKY(DATUM, JMENO)) jsou obsaženy sémantické n-tice (VYTVOR), (SCHUZKY), (DATUM), (JMENO), (VYTVOR, SCHUZKY), (SCHUZKY, DATUM), (SCHUZKY, JMENO), (VYTVOR, SCHUZKY, DATUM) a (VYTVOR, SCHUZKY, JMENO).
- *Sémantický strom* - hiararchická závislost mezi jednotlivými sémantickými koncepty a slovy vstupní promluvy [5, s. 11]. Příkladem sémantického stromu může být třeba VYTVOR(SCHUZKY(DATUM(zítra), JMENO(markéta)))
- *N-gram* - souvislá sekvence n slov z dané posloupnosti, textu či promluvy.
- *Trénovací množina* - prvky jsou dvojice (trénovací příklad, cílová třída). Je použita pro trénování modelu HDM.

- *Racionální jádrové funkce* - komplexnější struktura pro reprezentaci trénovací množiny. Funkce je založená na konečných transducerech, což umožňuje vyčíslení jádrové funkce mezi dvěma konečnými automaty (akceptory či transducery) [5].

- *Příznakový vektor* - je složen z původního slova a lingvistických příznaků [7, s. 74]. Tyto příznaky jsou generovány ze vstupní promluvy (např. předpokládané četnosti slovních n-gramů $n = 1, 2...n_{max}$) [6].

- *Terminální symbol* - znak nacházející se v abecedě.

- *Neterminální symbol* - znak nenacházející se v abecedě.

- *Stochaistická bezkontextová gramatika* - generuje bezkontextový jazyk, značí se G, a odpovídá čtveřici $G = (\mathcal{N}, \Sigma, \mathcal{R}, S)$ [8], kde $\mathcal{N}$ je množina neterminálních symbolů, Σ je množina terminálních symbolů, $\mathcal{R}$ je množina pravidel, ve tvaru $A \rightarrow \beta$,kde $A \in \mathcal{N}$ a β je řetězec libovolné délky složený z terminálních a neterminálních symbolů a $S \in \Sigma$ je startovací symbol.

- *SVM (Support Vector Machines) klasifikátor* - jeho princip je založený na podpůrných vektorech [9], [10] a [5]

- *Klasifikátory sémantických entit* - referenční diskriminativní model využívající SVM klasifikátoru pro predikci sémantického stromu.

- *Binární klasifikátor* - predikuje přítomnost dané n-tice v sémantickém stromu odpovídající vstupní, neznámé promluvě [5, 15].

- *Parser se skrytým vektorovým stavem* - referenční model, vůči kterému je porovnáván přínos HDM [5, s. 13].

3.1 Vstupní vrstva

Model HDM dokáže zpracovat vstup v podobě slovní nebo fonémové mřížky či řetězce, ze které(ho) je vypočítán příznakový vektor. Z důvodu výpočetní náročnosti toho procesu je použita teorii racionálních jádrových funkcí. Jednou z výhod této teorie je možnost přímého výpočtu hodnot racionálních jádrových funkcí $K(u_k, u_j)$ mezi dvěma promluvami $u_k, u_j \in T$ z trénovací množiny.

Před použitím teorie racionálních jádrových funkcí musí být nejprve trénovací množina zpracována do podoby minimálního deterministického WFST[4], který je použit za účelem rychlého výpočtu racionální jádrové funkce. Výhodou použití této formy indexace je možnost kompilace libovolného uživatelského dotazu ve formě řetězce nebo mřížky do podoby konečného automatu [12].

K normalizaci hodnot jádrových funkcí do intervalu $[0, 1]$, je použita normalize konkrétněji popsaná v [13]. Normalizované hodnoty jádrových funkcí jsou přímo použity v SVM klasifikátorech skryté vrstvy. Použití WFST k reprezentaci jednotlivých promluv nám umožňuje trénovat a porozumět modelu, který bere v úvahu nejistoty ASR[5] mřížky. Pro implementaci vstupní vrstvy byla použita knihovna OpenFST [14], která zajišťuje efektivní implementaci WFST algoritmů.

Vstupní vrstva- vypočítání hodnot jádrových funkcí.	
Vstup:	**Výstup:**
Slovní nebo fonémová mřížka. Slovní nebo fonémový řetězec.	Normalizované hodnoty jádrových funkcí.

[4] *Weighted Finite State Transducers*, vážený konečný stavový transducer[11]
[5] *Automatic Speech Recognition*, automatické rozpoznání mluveného jazyka

3.2 Skrytá vrstva

Ve skryté vrstvě je použit již zmíněný model STC sloužící ke zmenšení vstupního prostoru z několika tisíc možností na menší sadu sémantických n-tic. Použitý model obsahuje několik modifikací. Narozdíl od původního modelu používá sadu binárních klasifikátorů natrénovaných k predikci přítomnosti či absence sémantických n-tic. Volba délky sémantických n-tic je volitelná. Při volbě kratších n-tic bylo dle experimentů v [5] dosaženo robusnější klasifikace, ale původní model STC měl s volbou krátkých n-tic problém.

Ve skryté vrstvě HDM je původní výstup STC modelu nahrazen parsovacím algoritmem. Rekonstrukce sémantického stromu je založena na trénování klasifikátorů z anotovaných dat. Pro každou sémantickou n-tici je použit jeden klasifikátor. U málo četných n-tic je přesnost klasifikace nízká [5], proto je pro trénování skryté vrstvy použita množina obsahující všechny sémantické n-tice vyskytující se v tréningových datech více než N-krát.

Při dekódování je ve skryté vrstvě složené z binárních SVM klasifikátorů získaných nad trénovací množinou použita transformace dosud neviděné promluvy na příznakový vektor. Pro predikci neviděné promluvy je nutné vyčíslit hodnoty racionálních jádrových funkcí mezi neviděnou promluvou a promluvami z trénovací množiny, a tyto hodnoty využít při klasifikaci n-tic ve výstupním stromě. Hodnota příznakového vektoru odpovídá vzdálenosti k rozhodovací nadrovině klasifikátoru detekující přítomnost sémantické n-tice.

Skrytá vrstva- předzpracování jádrových funkcí modelem STC.

Vstup:	**Výstup:**
Normalizované hodnoty jádrových funkcí.	Příznakový vektor reprezentující vzdálenosti sémantických n-tic k rozhodovací nadrovině SVM.

3.3 Výstupní vrstva

Výstupní vrstva používá sémantickou gramatiku, která je podobná bezkontextové gramatice s několika málo rozdíly. Oproti bezkontextové gramatice nejsou symboly gramatiky děleny na terminální a neterminální, existuje pouze jediná množina sémantických konceptů. Důsledkem chybějících terminálních symbolů jsou pravidla pro reprezentaci slov pozměněna. Sémantická gramatika se skládá ze sady sémantických konceptů, gramatických pravidel závislých na promluvě a počátečního symbolu parsovacího algoritmu. Ze získané gramatiky je generován nejpravděpodobnější sémantický strom. Tuto pravděpodobnost lze odhadnout za pomoci metod strojového učení. Generované stromy nejsou seřazeny.

Pro posteriorní odhad pravděpodobnosti výskytu konceptu s danými potomky v sémantickém stromě pro danou vstupní promluvu je použit diskriminativní model obsahující SVM klasifikátoru s RBF[6] jádrem schopný klasifikace do více tříd. Tento klasifikátor je použit pro získání pravděpodobnostního rozdělení nad cílovou třídou. K natrénování klasifikátoru je potřeba anotovaný sémantický strom, který pro každý klasifikátor předpovídá další rozšíření konceptu a transformuje ho do cílové třídy. Tento postup zajistí, že je výstupní klasifikátor natrénován za použití příznakového vektoru a cílové třídy pro každý koncept. Po natrénování klasifikátorů je možné predikovat chtěné posteriorní odhady.

Za pomoci klasifikátoru schopného klasifikace do více tříd je pro každý koncept předpovězeno pravděpodobností rozdělení přes odpovídající sady množin všech možných následovníků. Sada obecných pravidel R je rozšířena získanými pravděpodobnostmi a je vytvořena nová sémantická gramatika, ve které je sada obecných pravidel nahrazena nově vytvořenou množinou R_u. Pro získání nejpravděpodobnějšího abstraktního sémantického stromu ze nově vytvořené sémantické gramatiky vstupní promluvy je použit algoritmus určení abstraktního sémantického stromu[5, s. 76], využívající prohledávání s nejmenší cenou (BFS[7]).

[6] *Radiální bázová funkce* [5, s. 28] a [15]

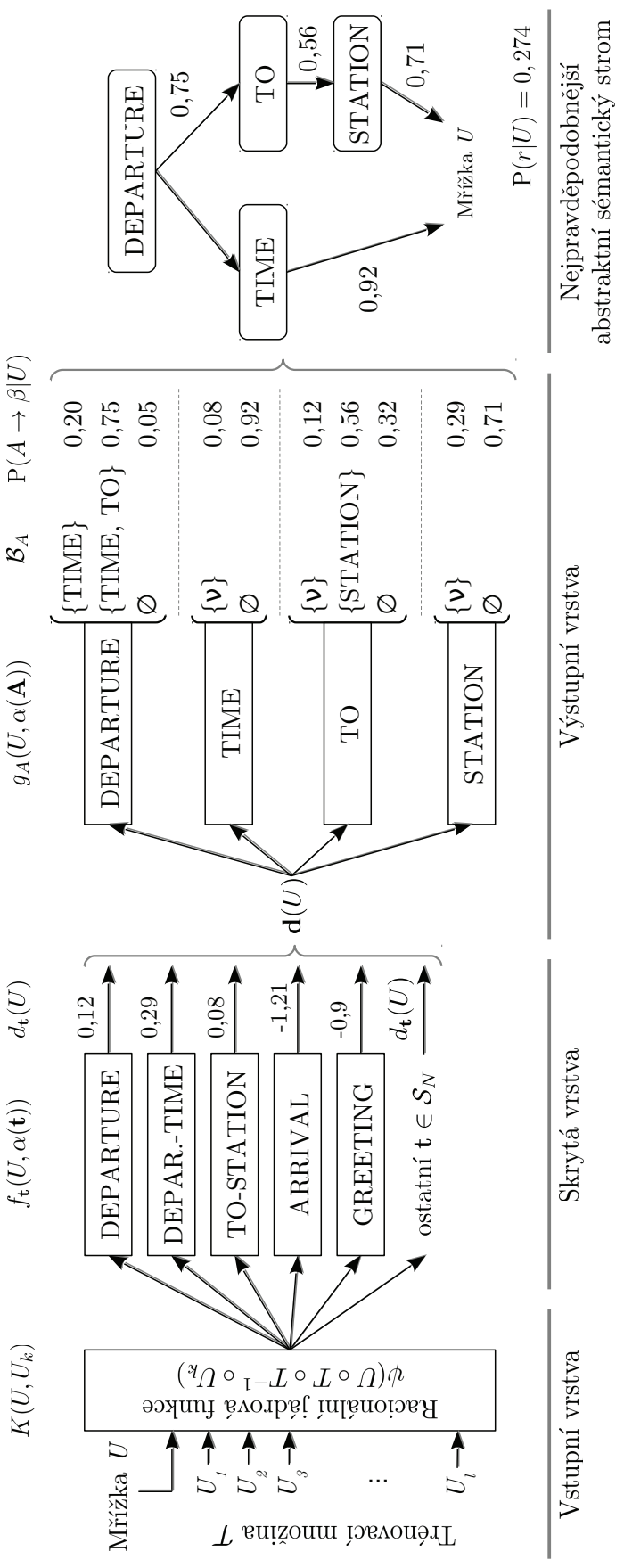
[7] *Best First Search*, algoritmus pro prohledávání stavového prostoru, který postupuje od vrcholu prohledávání vybere ten uzel, který je nejvíce slibný vzhledem k specifikovanému pravidlu

Algoritmus iterativně expanduje uzly sémantického stromu od kořenového symbolu s_o s využitím pravidel z R_u. Výsledný parciální strom r je reprezentován jako seznam použitých pravidel. Tento algoritmus také umožňuje nalezení n-tého nejpravděpodobnějšího stromu pouhým pokračováním BFS algoritmu po nalezení prvního sémantického stromu. V rekurzivních pravidlech je zakázáno, aby se v množině všech možných následovníků vyskytoval koncept, pro který je tato množina následovníků vytvořena. Tato podmínka brání ve vytvoření nekonečného algoritmu, ale není pro algoritmus limitující.

Výstupní vrstva- slouží k nalezení n-nejlepších sémantických stromů.	
Vstup:	**Výstup:**
Příznakový vektor reprezentující vzdálenosti sémantických n-tic k rozhodovací nadrovině SVM.	Nejpravděpodobnější sémantický strom. *N-nejlepších* sémantických stromů.

3.4 Shrnutí

Model HDM inovuje původní STC model a snaží se o eliminaci jeho nedostatků. Po natrénování modelu je možné získat ze slovní nebo fonémové mřížky (řetězce) nejpravděpodobnější sémantický strom případně n-nejlepších sémantických stromů. Tento model je zároveň za pomoci testovacích dat schopný posoudit přesnost rozpoznání sémantických stromů formou konceptové přesnosti *cAcc*. Na obr. 3.1 je uvedeno schéma tohoto modelu.



Obrázek 3.1: Schéma hierarchického diskriminarivního modelu.

Kapitola 4

Strategie výběru dat

Tato kapitola je zaměřená na popis strategií a jejich algoritmů, navržených k výběru dat pro trénování hierarchického diskriminativního modelu. Pro trénování modelu jsou navrhnuty čtyři metody snažící se o prozkoumání vlivu výběru vstupních trénovacích dat na průběh simulace modelu:

- náhodný výběr různých typů trénovacích dat
- výběr dat dle jejich neurčitosti
- výběr dat dle špatné predikce
- výběr dat dle slovníku

Experimentální ověření navržených metod se nachází v kapitole 5. Jedním z cílů této diplomové práce bylo provést plnou automatizaci procesu výběru dat. Pro tento účel byl zvolen programovací jazyk *Python*. U popisu jednotlivých strategií jsou zmíněny i použité funkce, které byly potřeba k naprogramování a automatizaci jejich postupu.

Pro upřesnění různých druhů dat a parametrů vyskytujících se v této práci je v kapitolách: *„Specifikace použitých dat“* a *„Specifikace použitých pojmů“* uveden jejich přehled s krátkým teoretickým vysvětlením. Pro experimentální ověření metod zabývající výběrem dat dle hodnoty jejich *posteriorní pravděpodobnosti* a výběrem dat dle hodnoty *F-skóre* je použit stejný postup pouze s odlišností výběrového faktoru dat, proto je jejich strategie popsána v jedné kapitole nazvané: *„Strategie výběru dat dle predikovaných hodnot“*.

Pro model HDM byl použit řečový korpus TIA, který obsahuje celkem 10196 vět. Tento korpus byl rozdělen v poměru 72:8:20 na trénovací, heldout a testovací sady dat. Vstupem jednotlivých strategií je množina uskutečněných dialogů určená pro trénování modelu neboli testovací sada. Z této sady jsou vybrána data jednou z uvedených strategií, kterou jsou ohodnocena jako nejvíce vhodná k anotaci. Tato vhodnost je dána hodnotou výběrového faktoru: *pravděpodobnost promluv*, *V-míra* či *F-score* dle zvolené strategie. Vybraná data jsou použita pro natrénování modelu HDM. K optimalizaci parametrů modelu byla použita heldout sada dat. Na vstup natrénovaného modelu jsou poté přivedena testovací data. Po skončení běhu modelu je získána konceptová přesnost *cAcc*, která je důležitá pro hodnocení přesnosti predikce sémantických stromů tvořených sémantickými koncepty. Tato přesnost je použita pro vyhodnocení jednotlivých strategií a její přesnější popis lze nalézt v kapitole: „Specifikace použitých pojmů“.

4.1 Schéma výběru vhodných trénovacích dat:

Navržené strategie výběru dat určených pro trénování HDM modelu jsou součástí níže popsaného základního schématu všech experimentů. Cílem návrhu metod je vhodným výběrem trénovacích dat dosáhnout zvýšení konceptové přesnosti *cAcc*.

- Pro použití schématu je potřeba model porozumění mluvené řeči (HDM) trénovaný z malého množství dat (10%)
- Vstupem je velké množství neanotovaný dat
- Cílem je vybrat nejvhodnější data (dalších 10%, 20%, 40%) pro oanotování a přetrénování modelu
- Neanotované věty jsou ohodnoceny výběrovým faktorem (posteriorní pravděpodobnost, V_u, F-skóre), oříznuty a seřazeny (vzestupně, sestupně)
- Data jsou oanotována (jsou vybrány anotace, protože experiment je simulován)
- Model je natrénován a je vyhodnocena konceptová přesnost *cAcc*

4.2 Specifikace použitých dat

Již zmíněný korpus TIA, obsahuje dva druhy dat: manuální a reálná. V této práci je toto označení použito především pro strategii náhodného výběru vstupních dat. Pojem **manuální data** signalizuje, že data vznikla za pomoci anotátora. Data byla ručně vytvořena tak, aby systém na ně trénovaný vždy porozuměl správně. Sada manuálních dat obsahuje celkem 439 vět.

Druhou možností vstupních dat pro model HDM jsou **reálná data**, která byla získána dvěma způsoby. První část dat byla nahrána pomocí systému simulujícího chování budoucího dialogového systému za pomoci posloupnosti jednotlivých poddialogů. Druhá část dat byla získána cíleným sběrem promluv obsahující vybrané sémantické entity. Sada reálných dat obsahuje celkem 7639 vět.

4.3 Specifikace použitých pojmů

Během popisu strategií výběru trénovacích dat a jejich experimentálních ověření jsou často použity pojmy: *F-skóre*, *Míra* V_u a *pravděpodobnost promluv*, obecně nazvané jako výběrový faktor, proto je v této kapitole uvedena jejich bližší specifikace. Kromě těchto výběrových faktorů je zde definován parametr pro určení přesnosti výsledků, neboli konceptová přesnost *cAcc*.

4.3.1 Konceptová přesnost *cAcc*

Tato míra je použita v kapitole: „Experimentální ověření“, pro vyhodnocení přesnosti jednotlivých strategií. Konceptová přesnost slouží k vyhodnocení přesnosti predikce sémantických stromů tvořených sémantickými koncepty a je schopna podchytit míru shody mezi dvěma různými sémantickými stromy. Tuto míru *cAcc* je lze definovat jako:

$$cAcc = \frac{N - D - S - I}{N} \tag{4.1}$$

kde N odpovídá počtu referenčních konceptů, H počtu správných konceptů, S počtu chyb substituce, D počtu chyb vynechání a I počtu chyb vložení [5, s. 111].

4.3.2 Pravděpodobnost promluv

Na výstupu HDM modelu je získáno n-nejlepších sémantických stromů $r_1, r_2...r_n$ s přiřazenými pravděpodobnostmi $P(C = r_i \mid U = u)$, kde C je absolutní sémantický strom, C je promluva, pozorovaná proměnná a u je vstupní promluva. Pravděpodobnost částečného sémantického stromu $P(r \mid u)$, lze vyčíslit i pro plně neexpandované sémantické stromy. Nelze je tedy přímo použít k přiřazení aposteriorní pravděpodobnosti, proto je použita aproximace:

$$P(C = r_i \mid U = u) \approx \frac{P(r_i \mid u)}{\Sigma_{k=1}^{n} P(r_k \mid u)} \tag{4.2}$$

Tato aproximace zajišťuje normalizaci přiřazených aproximací tak, aby platilo:

$$\Sigma_{i=1}^{n} P(C = r_i | U = u) = 1 \tag{4.3}$$

4.3.3 Míra V_u

Tato míra je použita pro výběr dat v metodě: „Strategie výběru dat dle jejich slovníku“. Postup pro získání hodnoty míry V_u zahrnuje získání prvotního slovníku $\mathcal{V}$ z předepsaných 10% trénovacích dat. Slovník lze definovat jako množinu jednotek, v tomto případě slov. Nyní pro jakoukoliv promluvu V_i, která se nenachází v původních 10% dat lze získat hodnotu míry V_u definovanou jako:

$$V_u = \frac{|\mathcal{V} \cap V_i|}{V_i} \tag{4.4}$$

Míra V_u se pohybuje v rozsahu $\langle 0, 1 \rangle$ a její hodnota udává, kolik slov z hodnocené promluvy se nachází ve slovníku $\mathcal{V}$. Míru V_u s hodnotou **1** znamená, že všechna slova již jsou ve slovníku $\mathcal{V}$. Naopak hodnota **0** indikuje, že žádné ze slov promluvy se ve slovníku nenachází.

4.3.4 F-skóre

Pro objektivní vyhodnocení výstupů modelu porozumění mluvené řeči je možné kromě konceptové přesnosti *cAcc* definovat tzv. F-skóre. Tato míra je použita pro výběr dat v metodě: *„Strategie výběru dat dle predikovaných hodnot“*. K definici tohoto parametru je zapotřebí znát hodnoty *míry úplnosti R* (recall) a *přesnost P* (precision):

$$R = \frac{TP}{TP + FN} \tag{4.5}$$

$$P = \frac{TP}{TP + FP} \tag{4.6}$$

kde *TP* je počet správně predikovaných výskytů daného jevu, *FP* je počet chyb, kdy byl jev predikován a *FN* je počet chyb, kdy jev nebyl predikován, ale referenční data je obsahují. *F-skóre* je pak definováno jako harmonický průměr měr P a R:

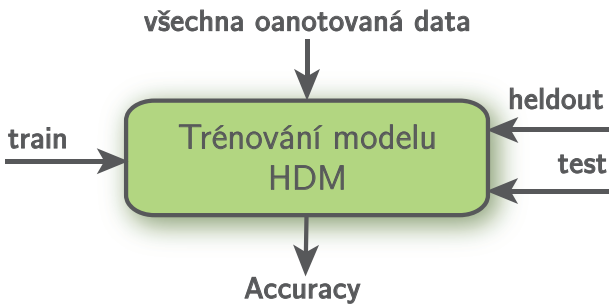
$$F = 2\frac{P \cdot R}{P + R} \tag{4.7}$$

Koncept	*F-skóre*	*přesnost P*	*míra úplnosti R*
DIKY	96,2	100,0	92,68
SPOJ	93,33	98,25	88,89
SOUHLAS	90,91	95,89	86,42
KONFERENCE	88,89	100,0	80,00
VYTVOR	86,08	85,00	87,18
INTERVAL	84,85	92,11	78,65
NE	84,30	94,44	76,12
HELLO	82,76	97,30	72,00
JMENO	81,93	90,27	75,00
ZJISTI	81,67	87,50	76,56
⋮	⋮	⋮	⋮

Tabulka 4.1: Ukázka hodnot F-skóre pro 10% předepsaných dat.

4.4 Strategie náhodného výběru dat

Pro náhodný výběr trénovacích dat je použita strategie zabývající se generováním pseudonáhodných čísel pomocí funkce **random.sample***(population,k)*. Díky této funkci bylo možné vybrat určité množství dat *k* ze seznamu *population*. V našem případě množství *k* odpovídá počtu vybraných vět ze zvolené sady dat (manuální, reálná, kombinace manuální a reálné). Například sada reálných dat obsahuje celkem 7639 vět, pro 10% dat by hodnota *k* byla rovna 764. Seznam *population* odpovídá seznamu identifikačních čísel vět pro celou sadu. Ověření této strategie je provedeno pro dva druhy dat a jejich kombinaci. Více informací je možné nalézt v kapitole: *„Experimentální ověření“*. Ze zvolené sady dat jsou náhodně vybrána data v rozsahu 5 až 100%. Množství vybíraných dat závisí na zvoleném druhu: *manuální*, *reálná* či *kombinace manuálních a reálných dat*, více o těchto datech v kapitole: *„Specifikace použitých dat“*. Vybraná data jsou přiváděna na vstup modelu a je jimi trénován, jsou proto nazývány trénovacími.



Obrázek 4.1: Proces trénování.

Na obr. 4.1 jsou tato data označena jako **train**. Dalšími vstupy modelu jsou data **heldout** sloužící k optimalizaci parametrů modelu HDM a data **test** k otestování modelu. V takto pojmenovaných sadách dat jsou uvedeny pouze seznamy identifikačních čísel jednotlivých promluv. Proto je nutné do modelu jako další vstup použít **všechna oanotovaná data**. Model si pak na základě identifikačního čísla dokáže nalézt v této databázi odpovídající promluvu.

Po výběru dat pro trénování je kvůli výpočetní náročnosti modelu k jeho spuštění využito externí výpočetní středisko *METAcentrum*[8]. K automatizaci strategie je nutné vytvořit konfigurace a spouštěcí skripty pro každý výběr dat. Tyto skripty zajišťují správné spuštění úlohy a po jejich vytvoření se automaticky připojí k *METAcentru* a spustí se. Posledním krokem je vyhodnocení strategie za pomoci grafu závislosti konceptové přesnosti na množství vybíraných dat.

[8] *METAcentrum*, toto středisko je jednou z virtuálních organizací české Národní Gridové Iniciativy.

Vstupní požadavky: Přístup k modelu HDM v datovém úložišti. *METAcentrum*	
Vstup:	**Výstup:**
Soubor *manualni.scp*.	Graf zobrazující závislost parametru *cAcc* na množství
Soubor *train.scp*.	vstupních dat.

Algoritmus pro strategii náhodného výběru trénovacích dat:

1. *Vytvoř potřebné vstupní soubory.* V závislosti na zvoleném experimentu mohou nastat následující možnosti:

   - Spoj soubory *manualni.scp* a *train.scp* do jednoho seznamu. Vyber z něho pomocí funkce **random.sample** příslušný počet trénovacích dat.
   - Vyber data pouze ze souboru *manualni.scp* nebo *train.scp*.

2. *Vytvoř konfigurace.* Vytvoření konfigurací nutných pro spuštění úlohy. Kromě informacích nutných k vstupnímu spuštění modelu HDM obsahují konfigurace informace o cestách k vstupním souborům, název a adresář pro výstupní soubory.

3. *Vytvoř spouštěcí skripty.* Spouštěcí skript slouží k nastavení parametrů pro spuštění úlohy v *METAcentru*. Obsahuje například parametry jako jsou:

   - velikost využité paměti *mem=2gb*
   - maximální doba spuštění úlohy před jejím nedobrovolným ukončením *walltime=24h*
   - cesta ke konfiguraci vytvořené v minulém kroku

4. *Spusť skripty.* Po vytvoření potřebných souborů jsou skripty spuštěny. K tomu je potřeba se připojit se k *ssh* serveru pomocí pythonské funkce **ssh.connect**.

5. *Vyhodnoť získaná data.* Získaná data jsou uložena do adresáře specifikovaného v dříve vytvořené konfiguraci. Z vystupních souborů je vybrána výsledná konceptová přesnost *cAcc* a je zobrazena v grafu v závislosti na množství vybíraných dat.

Výsledky experimentů této strategie jsou uvedeny v kapitole 5.1.

4.5 Strategie výběru dat dle predikovaných hodnot

Pro první fázi trénování je vybráno předepsaných 10% trénovacích dat ze souboru *10_SCORE_train.scp*. Jsou označena jako *předepsaná*, aby bylo patrné, že i když tato data mohla být vybírána náhodně, byla použita data již vybraná za pomoci strategie náhodného výběru dat. Stejná data použita i pro strategii výběru dat dle slovníku. Volba stejných prvních 10% dat je zvolena za účelem možnosti kvalitnějšího porovnání jednotlivých metod. Takto jsou zajištěny pro všechny strategie stejné počáteční podmínky.

Modelem HDM bylo poté zpracováno zbylých 90% *neanotovaných* dat. Ze souboru s názvem *test.Ydec. ... .nbest.py* bylo pro každou promluvu možné nalézt *posteriorní pravděpodobnost* či hodnotu *F-skóre*. Z těchto neanotovaných dat bylo pak dále vybráno 10%, 20% a 40% jako nejvíce vhodných k anotaci dle hodnoty výběrového faktoru:

- **Posteriorní pravděpodobnost** - pro každou promluvu byla nalezena odpovídající hodnota posteriorní pravděpodobnosti a dle probíhajícího experimentu byla data seřazena a oříznuta

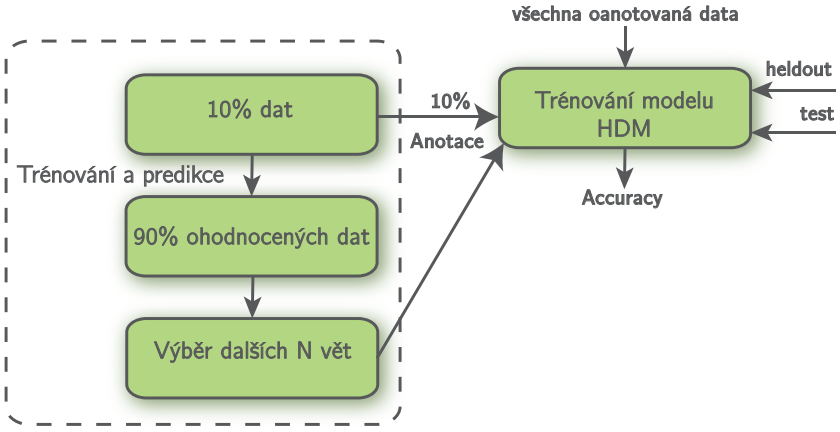
- **F-skóre** - pro každou promluvu byl nalezen nejpravděpodobnější sémantický strom skládající se z jednoho nebo více konceptů. Ze souboru *heldout.Ydec. ... .SCORE.cdc* byla získána tabulka s hodnotou *F-skóre* pro jednotlivé koncepty. Na základě této tabulky 4.1 lze dle obsažených konceptů vypočítat hodnotu F-skóre pro jednotlivé sémantické stromy za pomoci aritmetického průměru.

  Příklad: sémantický strom VYTVOR(KONFERENCE(JMENO)) by byl dle tabulky 4.1 ohodnocen jako: $(86,08 + 88,89 + 81,93)/3 = 85,63$. Pro promluvu by hodnota F-skóre byla $85,63$. Hodnoty F-skóre jsou přepočítány pro každou promluvu a dle ní jsou pak seřazeny (vzestupně, sestupně) a oříznuty, dle zvoleného experimentu.

Vybraná data byla dále použita k natrénování nového modelu za $10 + 10\%$, $10 + 20\%$ a $10 + 40\%$ dat. Výstupem modelu je konceptová přesnost *cAcc*, která je zobrazena do grafu v závislosti na vybíraném množství dat. Postup této strategie je popsán schématem zobrazeném na obr. 4.2.

Pro tyto metody je model HDM použit dvakrát, nejprve pro prvních předepsaných 10% dat, aby byli získány predikované hodnoty výběrových faktorů a později pro celý

blok trénovacích dat (10% původních + N nově vybraných vět). Do modelu jsou ještě jako další vstupy přivedena oanotovaná, heldout a testovací data, jejichž funkce byla již představena v kapitele: 4.4.



Obrázek 4.2: Obecné schéma strategie výběru dat dle neurčitosti, predikce.

Vstupní požadavky: Přístup k modelu HDM v datovém úložišti *METAcentrum*	
Vstup: Soubor *manualni.scp*. Soubor *train.scp*. Soubor *10_SCORE_train.scp*	**Výstup:** Graf zobrazující závislost parametru **cAcc** na množství vstupních dat.

Algoritmus pro strategii výberu trénovacích dat dle neurčitosti či predikce

1. *Vytvoř potřebné vstupní soubory.* V tomto kroku jsou použita předepsaná data s názvem *10_SCORE_train.scp* a zbylá data ze sady manuálních a reálných dat jsou uložena jako *90_SCORE_test*.

2. *Vytvoř konfigurace, vytvoř spouštěcí skripty, spusť skripty.* Pro úlohu je vytvořena konfigurace a jí odpovídající spouštěcí skript. Takto vytvořené soubory jsou připraveny pro zpracování modelem HDM. Posledním krokem je proto spuštění skriptů.

3. *Získej potřebná data.* K získání predikovaných hodnot výběrových faktorů je využit soubor *test.Ydec.tia. ... .nbest.py*, ze kterého jsou získány hodnoty posteriorních pravděpodobností nebo nejpravděpodobnější sémantické stromy, pro které je vypočítána aritmetickým průměrem hodnota *F-skóre* dle tabulky 4.1 získaná se souboru *heldout.Ydec. ... .SCORE.cdc*.

4. *Vyber data.* Získané hodnoty jsou využity pro jednotlivé experimenty. Je možné data seřadit, či vybírat jen určitou část, záleží na druhu experimentu. Po úpravě dat je vybráno 10 až 90% dat. K natrénování modelu jsou použita za $10 + 10\%$, $10 + 20\%$, $10 + 40\%$ dat.

5. *Vytvoř konfigurace, vytvoř spouštěcí skripty, spusť skripty a vyhodnoť získaná data.* Závěrečný postup probíhá obdobně jako u strategie popsané v kapitole 4.4.

Výsledky experimentů jsou uvedeny v kapitolách 5.1 a 5.3.

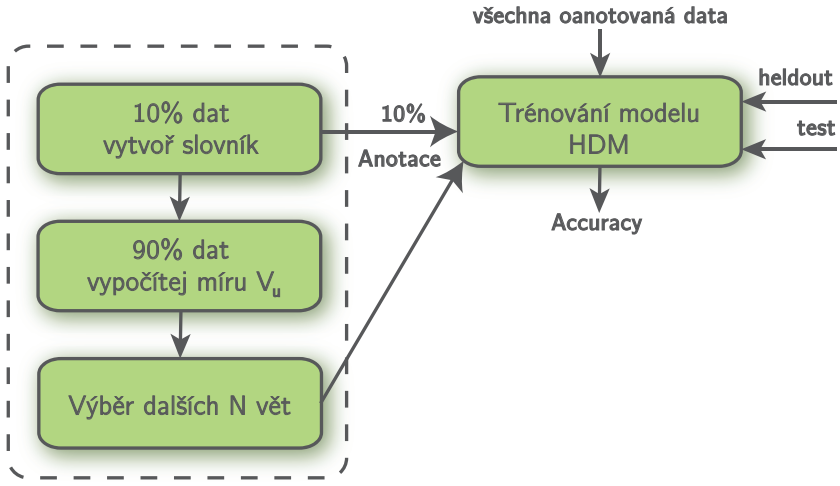
Poznámka: Algoritmus lze drobnou úpravou předělat tak, aby si předepsaných 10% dat vytvářel sám.

4.6 Strategie výběru dat dle jejich slovníku

Oproti strategii 4.5 není potřeba prvotní získání predikovaných hodnot výběrového faktoru. Pro výběr dat dle jejich slovníku je jeho hodnota vypočítána pouze za pomoci předepsaných 10% dat nacházejících se v souboru *10_SCORE_train*. Z těchto dat je vytvořen slovník $\mathcal{V}$, pomocí něhož je pro každou promluvu nacházející se v kombinaci manuálních a reálných dat (bez předepsaných 10%) vypočítána hodnota V_u, blíže popsaná v kapitole: 4.3.3.

Neanotovaná data s vypočtenou hodnotou V_u byla pro odpovídající experiment seřazena a následně v závislosti na probíhajícím experimentu, mohla být z této množiny dat vybrána jen určitá část (oříznutí prvních 25% dat). Nakonec byla dále použita k natrénování modelu HDM za $10 + 10\%$, $10 + 20\%$ a $10 + 40\%$ dat.

Obdobně jako v předešlých strategií byla vybraná data vstupem modelu HDM spolu s **heldout**, **test** a **všemi anotovanými daty**, jejichž funkce již byla představena v kapitole: 4.4. Výstupem modelu je konceptová přesnost zobrazená v závislosti na vybíraném množstvím dat formou sloupcového grafu. Postup této strategie je popsán následujícím schématem:



Obrázek 4.3: Obecné schéma strategie výběru dat dle jejich slovníku.

Vstupní požadavky:	Přístup k modelu HDM v datovém úložišti *METAcentrum*.
Vstup: Soubor *manualni.scp*. Soubor *train.scp*. Soubor *10_SCORE_train.scp*	**Výstup:** Graf zobrazující závislost parametru *cAcc* na množství vstupních dat.

Algoritmus pro strategii výběru dat dle jejich slovníku:

1. *Vytvoř potřebné vstupní soubory.* V tomto kroku jsou použita předepsaná data s názvem *10_SCORE_train.scp*, ze kterých je vytvořen slovník $\mathcal{V}$.

2. *Vypočítej míru* V_u. Pro data ze souboru kombinující manuální a reálná data, která se nenacházející v *10_SCORE_train.scp* (zbylých 90%) a pro každou promluvu v nich obsaženou vypočítej V_u.

3. *Vyber data.* Dle zvoleného experimentu je možné vybírat ze seřazených dat, či vybírat jen jejich určitou část. Po výběru data přidej k předepsaným 10% a vytvoř soubory trénovacích dat o velikosti $10 + 10\%$, $10 + 20\%$ a $10 + 30\%$ dat.

4. *Vytvoř konfigurace, vytvoř spouštěcí skripty, spusť skripty a vyhodnoť získaná data.* Závěrečný postup probíhá obdobně jako u strategie popsané v kapitole 4.4.

Výsledky experimentů jsou uvedeny v kapitole 5.4 .

Kapitola 5

Experimentální ověření

Následující kapitola je zaměřená na experimentální ověření metod blíže specifikovaných v kapitole: *„Strategie výběru dat“*. Metody byly testovány nad sémantickým korpusem TIA a vyhodnoceny za pomoci modelu HDM. Hlavním cílem provádění experimentů bylo zjištění závislosti nárůstu konceptové přesnosti získané z modulu porozumění mluvené řeči na zvolené strategii výběru dat pro trénování modelu. K výběru vstupních dat se přistupovalo pomocí čtyř strategií. První z nich využívá zcela **náhodný výběr trénovacích dat** a pro ni uskutečněné experimenty jsou popsány v kapitole: *„Náhodný výběr trénovacích dat“*. Tyto experimenty jsou rozděleny do tří kategorií:

- využití manuálních dat
- využití reálných dat
- použití kombinace manuálních a reálných dat

Nejlepších výsledků po experimentálním ověření dosahovala data vytvořená kombinací manuálních a reálných dat. Tento výsledek vedl k použití tohoto druhu vstupních dat v experimentálních ověření zbylých tří strategií.

Použitím náhodnosti vstupních dat u první metody bylo dosaženo odhadu středních hodnot konceptové přesnosti *cAcc* pro dané množství vybíraných dat. Díky tomuto odhadu je možné sledovat zlepšení či zhoršení výsledků zbylých strategií. Tento výsledek vedl k jejich využívání ve výsledných grafech ostatních strategií. Pro snazší orientaci jsou tyto výsledky označeny jako *baseline*.

Předpoklad: při vhodném výběru dat pro trénování modelu HDM (strategií 4.4 nebo 4.5), bude dosaženo vyšší hodnoty konceptové přesnosti než v případě *baseline*.

Pro strategii zabývající se **výběrem dat dle neurčitosti** je důležitý výběrový faktor dat, který odpovídá **pravděpodobnosti správného určení jednotlivých promluv**. Více informací o způsobu výpočtu této pravděpodobnosti lze nalézt v kapitole: *„Pravděpodobnost promluv“*. Experimentální ověření této metody je popsáno v kapitole: *„Výběr dat dle jejich neurčitosti“*. Po uskutečnění strategie je vytvořen soubor trénovacích dat, který je jedním ze vstupů hierarchického diskriminativního modelu. Jedním z výstupů modelu je vyhodnocení přesnosti systému realizované jako již dříve blíže popsaná míra *cAcc*.

Ve strategie zabývající se výběrem trénovacích dat **dle špatné predikce** jejíž experimentální ověření je popsáno v kapitole: *„Výběr dat dle špatné predikce“*, je použit stejný postup jako u strategie výběru dat dle jejich neurčitosti. Jediný rozdíl je ve výběrovém faktoru jednotlivých metod. Pro tuto metodu odpovídá tento faktor tzv. F-skóre, konkrétněji popsaný v kapitole: *„F-skóre“*.

Poslední strategie se zabývá **výběrem za pomoci slovníku jednotlivých promluv** a její experimenty jsou popsány v kapitole: *„Výběr dat dle slovníku“*, pro tuto metodu je použit výběrový faktor definovaný v kapitole: *„Míra V_u“*. Je předpokládáno, že výběrem dat za pomoci určitého faktoru výběru by v porovnání s náhodným výběrem trénovacích dat mělo být dosaženo vyšší hodnoty konceptové přesnosti. Pro zanalyzování strategií bylo provedeno více experimentů zabývajících se různým řazením a výběrem dat.

Všechny experimenty byly automatizovány za pomoci kódu vytvořeného v programovacím jazyku *Python*. Jako první byly (dle požadavků jednotlivých metod) vytvořeny soubory s trénovacími daty. K těmto vstupním souborům byly automaticky vytvořeny odpovídající konfigurace a spouštěcí skripty. Na závěr bylo provedeno spuštění skriptů a vyhodnocení metod.

5.1 Náhodný výběr trénovacích dat

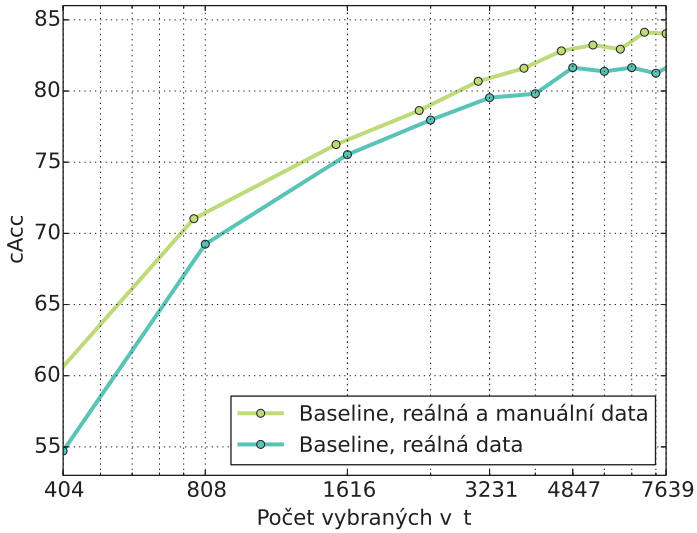
Ověření této strategie bylo dle použitých vstupních dat (*manuální*, *reálná* a *kombinace manuálních a reálných dat*) rozděleno na tři základní experimenty. V každému experimentu bylo vybráno dané množství vstupních dat použitých k trénování hierarchického diskriminativního modelu. Více informací o použitých datech v kapitole: *„Specifikace použitých dat“*. Po natrénování modelu bylo provedeno testování a ve výstupních souborech byla nalezena míra *cAcc*, sloužící k vyhodnocení přesnosti systému automatického rozpoznání řeči. V našem případě je tato míra využita k analýze použité strategie. Její přesnější definice uvedena v kapitole: *„Konceptová přesnost cAcc“*. Postup strategie těchto experimentů je popsán v kapitole: *„Strategie náhodného výběru dat“*.

Při výběru ze sady **manuálních dat** jsou experimenty provedeny pro 30% až 100% dat. U množství dat nižšího než 30% nemohly být experimenty provedeny z důvodu nesplnění požadavku minimálního množství dat potřebného pro natrénování modelu HDM. Po natrénování modelu je otestován a jsou získány výsledné hodnoty *cAcc*. Se zvyšujícím se počtem dat je možné pozorovat nepravidelnost hodnot konceptové přesnosti, která je způsobena náhodností vybíraných dat pro trénování modelu. Tato nepravidelnost je u použití manuálních dat viditelná zejména díky menší množině dat (439 vět), ze které jsou trénovací data vybírána. Ve výsledném grafu při použití reálných (7639 vět) a kombinace manuálních a reálných dat (8078 vět) 5.1 již tyto nepravidelnosti viditelné nejsou.

Výsledné hodnoty konceptové přesnosti při vybírání z množiny manuálních dat jsou velmi nízké, jedním z důvodu tohoto výsledku je malé množství trénovacích dat. Sada manuálních dat obsahuje celkem 439 vět, oproti tomu sada reálných dat 7639 vět. Výsledky získané po natrénování modelu není možné přímo porovnávat pro stejný procentuální výběr dat. Například při výběru 30% dat pro trénování modelu je ze sady manuálních dat vybráno 132 vět, zatímco ze sady reálných dat 2291 pro trénování. Jediná možnost pro porovnání výsledků při použití těchto dvou druhů dat je pro jejich přibližně stejné množství. Tento požadavek je splněn pro 90% manuálních (395 vět) a 5% reálných (381 vět), hodnota *cAcc* pro reálná data odpovídá 54% a pro manuální 27, 02%. Tento výsledek může být způsoben množstvím slov v jednotlivých promluvách a jejich přínosností pro trénování modelu. Za přínosná data lze označit taková data, která jsou pro model nová a zajistí modelu rozšíření jeho dosavadního slovníku. Výsledný rozsah *cAcc* pro manuální data se pohybuje v rozsahu 12% až 27%.

Další experiment této metody se zabývá výběrem z množiny **reálných dat**. Díky rozsáhlejší množině trénovacích dat je dosaženo plynulejšímu průběhu nárůstu konceptové přesnosti, než tomu bylo u manuálních dat. Hodnot *cAcc* pro 5% až 100% vybíraných dat se pohybuje v rozsahu 54% až 82%. Pro data vytvořená výběrem ze souboru obsahující, jak **manuální**, tak **reálná** data se výstupní míra *cAcc* pro stejné množství vybíraných dat pohybuje mezi 59% až 84%.

Při porovnání získaných hodnot konceptové přesnosti v grafu 5.1 nastává stejný problém jako u experimentu využívající výběr z manuálních dat. Z tohoto důvodu jsou výsledné grafy mírně posunuty, aby odpovídala osa grafu stejnému počtu dat, využitých pro trénování modelu. Vyznačené body grafu znázorňují hodnoty *cAcc* pro 5%, 10% až 100% vybíraných dat. Použitím dat vytvořených kombinací manuálních a reálných sad je dosaženo vyšší hodnoty *cAcc* pro jakékoliv množství vybíraných dat, než u výběru ze sady reálných dat. Tento výsledek vede k použití této kombinace dat pro zbylé metody za účelem co největšího zvýšení konceptové přesnosti *cAcc*.



Obrázek 5.1: Porovnání experimentů pro náhodný výběr dat.

Při náhodném výběru dat je dosaženo získání střední hodnoty konceptové přesnosti modelu HDM pro odpovídající množství vstupních dat. Tyto výsledky jsou proto využívány ve výsledných grafech ostatních strategií a jsou označeny jako *baseline*.

5.2 Výběr dat dle jejich neurčitosti

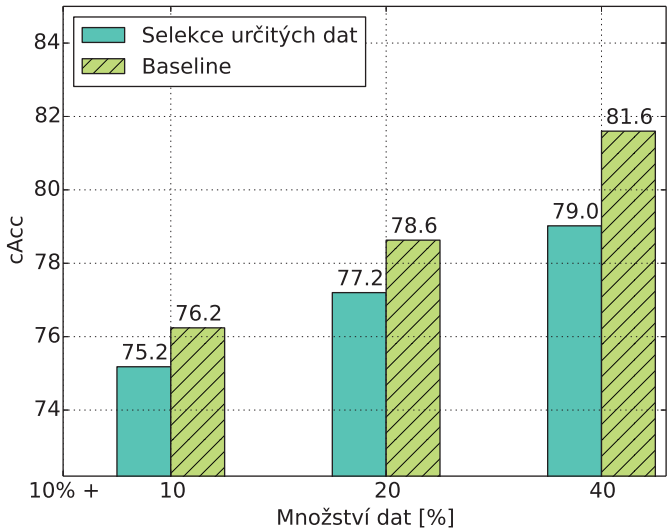
Další možnost výběru vstupních trénovacích dat je za pomoci pozorování predifikovaných posteriorní pravděpodobností. K použití metody je nejprve nutné natrénovat model z předepsaných 10% trénovacích dat. Tímto modelem bylo zpracováno zbylých 90% *neoanotovaných* dat. Na základě predifikovaných posteriorních pravděpodobností pak z těchto *neoanotovaných* dat bylo vybráno 10%, 20% a 40% jako nejvíce vhodných k anotaci. Tato data byla použita k natrénování modelu pro 10 + 10%, 10 + 20% a 10 + 40% dat. Strategie výběru dat je blíže popsaná v kapitole 4.5.

První experiment byl uskutečněn pro data seřazená od největší k nejmenší hodnotě získané posteriorní pravděpodobnosti. Jedná se o data, která model rozpoznal (dle posteriorní pravěpodobnosti) správně. Interval hodnot posteriorní pravděpodobnosti nejlepší hypotézy pro vybrané množství dat je možné vidět v tabulce 5.1:

10% +	10%	20%	40% dat
vybráno náhodně	0,99992 až 0,99935	0,99992 až 0,99648	0,99992 až 0,9623

Tabulka 5.1: Rozsah hodnot posteriorní pravděpodobnosti určitých dat.

Po spuštění experimentu pro výběr určitých dat, bylo trénováním modelu dosaženo výsledných hodnot konceptové přesnosti zobrazených v grafu 5.2:



Obrázek 5.2: Porovnání baseline a výběru určitých dat.

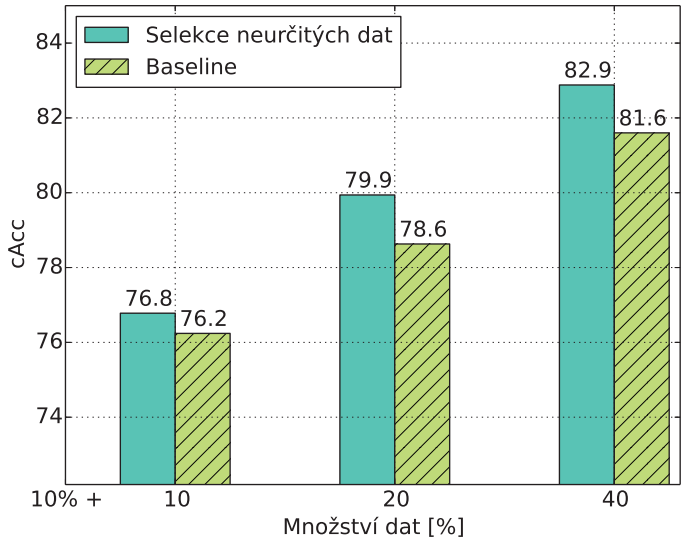
Z grafu 5.2 je pozorovatelné, že při použití dat, která jsou určitá je dosaženo nižších hodnot *cAcc*, než v případě *baseline*. Hodnota konceptové přesnosti se pohybuje od 75,18% až do 79,02%, tyto hodnoty jsou poměrně nízké. Na základě těchto výsledků lze předpokládat, že vybíraná data nejsou pro trénování přínosná, další experiment je proto zaměřen na výběr dat, která určitá nejsou. Tomuto předpokladu odpovídají data s nejnižší hodnotou posteriorní pravděpodobností.

Obdobně jako v předchozím případu byla data vybírána pro 10 + 10%, 10 + 20% a 10 + 40% dat. Rozsah hodnot pravděpodobnosti promluv pro vzestupně seřazená data jsou uvedeny v následující tabulce:

10% +	10%	20%	40% dat
vybráno náhodně	0,14761 až 0,54884	0,14762 až 0,72387	0,14761 až 0,85499

Tabulka 5.2: Rozsah hodnot pravděpodobnosti pro neurčitá data.

V grafu 5.3 je možné pozorovat pozvolné zvyšování rozdílu mezi *baseline* a strategií výběru neurčitých dat. Tento efekt je způsoben výběrem vhodných dat pro trénování modelu HDM. Se zvyšujícím se počtem dat se však množina vhodných dat snižuje, a tato křivka poté začíná stoupat pomaleji.



Obrázek 5.3: Porovnání baseline a výběru neurčitých dat.

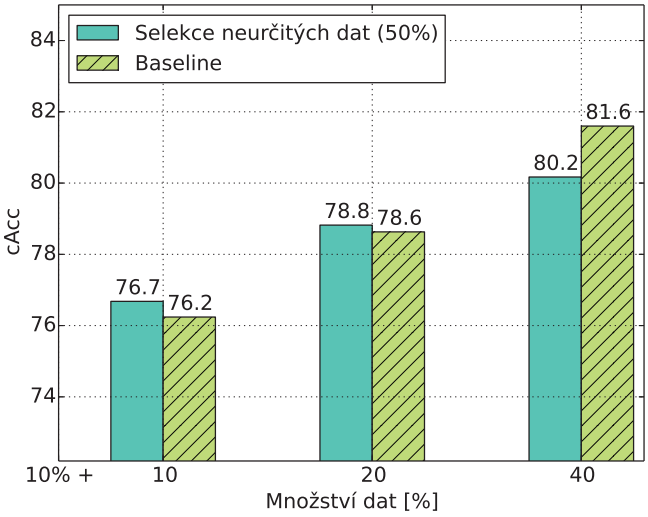
Výsledná hodnota *cAcc* se pohybuje v rozsahu 76,78% až 82,88%. Při vhodném výběru trénovacích (neurčitých) dat je dosaženo vyšší hodnoty konceptové přesnosti než při jejich náhodném výběru.

Rozdíl hodnot konceptové přesnosti mezi *baseline* a výběrem dat dle jejich neurčitosti, se postupně zvyšuje. Pro prvních 10 + 10% dat je tento rozdíl nejmenší, další experiment proto zkoumá možností zmenšení trénovací množiny o data, která nepřinášejí do učícího procesu přínosné informace. Nejprve je provedeno odstranění první poloviny vzestupně seřazených dat dle hodnoty posteriorní pravděpodobnosti, nový rozsah hodnot je zobr,l;azen v tabulce 5.3:

10% +	10%	20%	40% dat
vybráno náhodně	0,94748 až 0,97555	0,94748 až 0,99242	0,94748 až 0,99958

Tabulka 5.3: Rozsah hodnot pravděpodobnosti pro oříznutá data (50%).

Porovnáním grafu 5.3 s grafem 5.4 lze pozorovat, že odstraněním první poloviny seřazených dat určených pro trénovacích modelu je dosaženo nižší hodnoty konceptové přesnosti než v případě, kdy množina vstupních dat není oříznuta. Hodnota *cAcc* se po ořezu dat pohybuje v rozsahu 76,68% až 80,17%. Důvod dosažení nevhodných výsledků může být špatná volba prahu ořezu trénovacích dat.



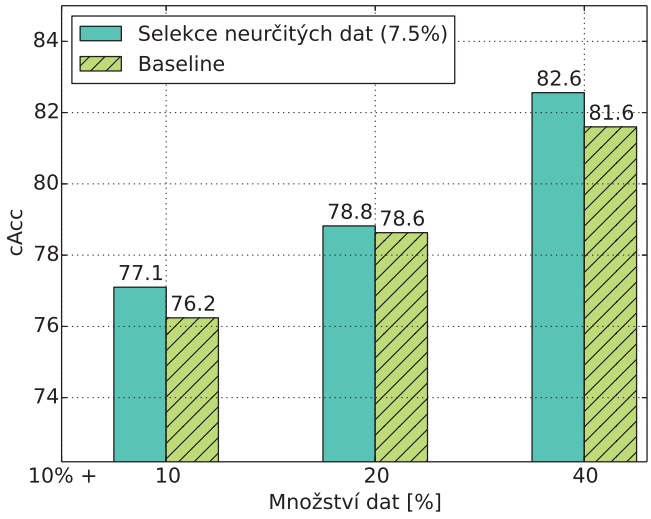
Obrázek 5.4: Výběr z oříznutých neurčitých dat (50%).

Jak je možné pozorovat v tabulkách 5.1 a 5.2 výsledná posteriorní pravděpodobnost jednotlivých sémantických stromů není rovnoměrně rozdělená. Nejvíce podobných hodnot se nachází v rozsahu $0,9$ až 1. Tento poznatek je viditelný i v grafu 5.3. Pro $10+10\%$ dat nenastává menší skok mezi hodnotami **cAcc** než pro $10+20\%$ vstupních trénovacích dat. Toto pozorování vede k zaměření dalšího experimentu na odstraněním dat s pravděpodobností správného určení sémantického stromu nižší než $0,5$. Zvolená hodnota odpovídá ořezu prvních $7,5\%$ trénovacích dat při vzestupném řazení dle posteriorní pravděpodobnosti. Nové hodnoty jsou zobrazeny v tabulce 5.3:

10% +	10%	20%	40% dat
vybráno náhodně	0,50807 až 0,68704	0,50807 až 0,82434	0,50807 až 0,95698

Tabulka 5.4: Rozsah hodnot pravděpodobnosti pro oříznutá data $(7,5\%)$.

V grafu 5.5 je možné pozorovat zvýšení konceptové přesnosti prováděného experimentu oproti výsledkům získaných hrubým ořezem první poloviny trénovacích dat v grafu 5.4. Pro prvních $10+10\%$ dat je dosaženo lepších výsledků než pro neoříznutá data. Hodnota *cAcc* se pohybuje mezi $77,10\%$ až $82,56\%$. Se zvyšujícím se počtem dat se v důsledku oříznutí rychleji vyčerpává trénovací množina obsahující vhodná data pro trénování. Z toho důvodu jsou pro $10+20\%$ a $10+40\%$ dat hodnoty *cAcc* nižší než u neoříznutých dat.



Obrázek 5.5: Výběr z oříznutých neurčitých dat $(7,5\%)$.

Pro přehled výsledků jednotlivých experimentů je zde uvedena tabulka 5.5, ve které je jako I_m označen průměrný nárůst konceptové přesnosti provedených experimentů oproti *baseline*.

	Hodnota *cAcc* pro			I_m [%]
	$10 + 10\%$	$10 + 20\%$	$10 + 40\%$ dat	
Baseline	76,24	78,63	81,60	x
Výběr určitých dat	75,18	77,20	79,02	-2,1
Výběr neurčitých dat	76,78	79,94	82,88	1,31
Výběr neurčitých dat (ořízntuto prvních 50% dat)	76,70	78.82	80,17	-0,3
Výběr neurčitých dat (ořízntuto prvních 7,5% dat)	77,10	78,82	82,56	0,81

Tabulka 5.5: Souhrnná tabulka výběru dat dle jejich neurčitosti.

Při výběru z množiny sestupně seřazených (určitých) dat bylo dosaženo výsledných hodnot konceptové přesnosti nižších než pro *baseline*. Při vzestupném seřazení dat (výběr z množiny neurčitých dat) jsou získány v porovnání s *baseline* vyšší hodnoty *cAcc*. Dalším ořezem vzestupně seřazených (neurčitých) dat je při špatné volbě prahové hodnoty dosaženo nižších hodnot konceptové přesnosti výsledků než bez použití prahové hodnoty.

V posledním experimentu zabývajícím se vhodnější volbou prahové hodnoty, která je zvolena jako $7,5\%$ je dosaženo pro menší počet vstupních trénovacích dat $(10 + 10\%)$ lepších výsledků. Při zvyšujícím se počtu dat začne docházet k vyčerpávání množiny vhodných dat pro trénování. Z tohoto důvodu jsou pro $10 + 20\%$ a $10 + 40\%$ získány nižší hodnoty konceptové přesnosti než pro experiment vybírající data bez prahové hodnoty.

Při výběru z množiny vzestupně seřazených (neurčitých) dat bylo v experimentech bez ořízntutí a s ořezem 7,5% dat dosaženo vyšších hodnot *cAcc* než v případě *baseline*. Pro tuto strategii výběru byl proto potvrzen předpoklad definovaný v úvodu této kapitoly (str. 26). Nejvyššího procentuálního nárůstu konceptové přesnosti v porovnání s baseline (1,31%) je dosaženo výběrem z množiny neurčitých dat bez jejího ořízntutí.

5.3 Výběr dat dle špatné predikce

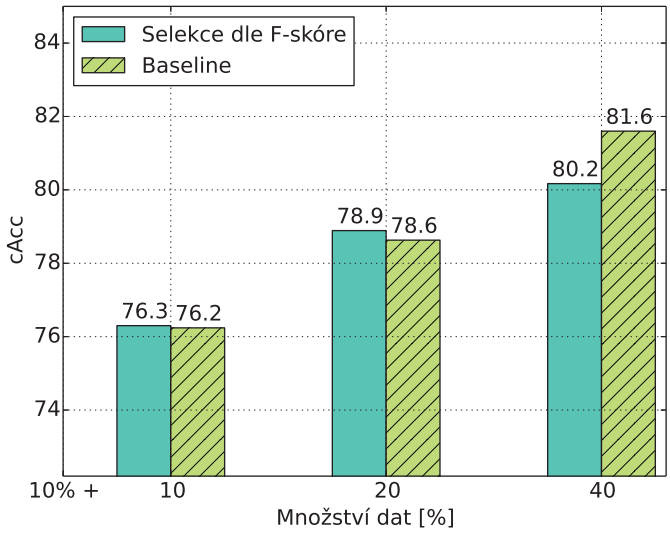
Jedním z dalších výběrových faktorů je tzv. *F-skóre*, neboli přesnost na úrovni jednotlivých konceptů, blíže popsaná v kapitole: *„F-skóre“*. Jako první byla zkoumaná možnost, ve které jsou vstupní data seřazena od největší k nejmenší hodnotě *F-skóre* vypočítaného pro sémantický strom každé promluvy. Jedná se o data, která byla vybrána na základě harmonického průběhu úplnosti a přesnosti jednotlivých konceptů.

Obdobně jako v jiných metodách byl natrénován model z přepsaných 10% trénovacích dat. Zbylých 90% dat bylo tímto modelem zpracováno. Na základě hodnoty F-skóre byla tato data vzestupně seřazena a z těchto *neanotovaných* seřazených dat bylo vybíráno 10%, 20% a 40% dat jako nejvíce vhodných k anotaci. Nakonec byla tato data pro 10 + 10%, 10 + 20% a 10 + 40% dat použita k natrénování nového modelu. Hodnoty *F-skóre* jsou zobrazeny v tabulce 5.6:

10% +	10%	20%	40% dat
vybráno náhodně	0,0 až 57,8	0,0 až 69,8	0,0 až 81,275

Tabulka 5.6: Rozsah hodnot F-skóre pro vzestupně seřazená data.

Po spuštění experimentu pro vzestupně seřazená data, vybírané na základě hodnoty *F-skóre*, byly získány následující hodnoty konceptové přesnosti zobrazené v grafu na 5.6:



Obrázek 5.6: Výběr dat dle F-skóre, data řazena vzestupně.

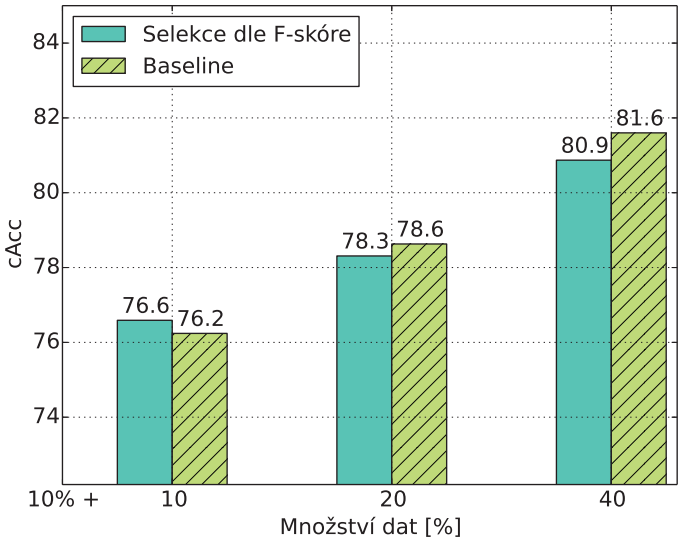
Při pozorování grafu na 5.6 je patrné, že hodnoty získané výběrem seřazených dat od nejmenší k největší hodnotě *F-skóre* pro $10 + 10\%$ a $10 + 20\%$, nejsou příliš odlišné od *baseline*. Lze tedy říci, že se tento způsob výběru vstupních trénovacích dat pro tento rozsah chová jako by byla data vybírána náhodně. Pro výběr $10 + 40\%$ je hodnota cAcc nižší než u *baseline*. Pro nás důležitá hodnota parametru **cAcc** se pohybuje v rozsahu 76,30 až 80,17%.

Z výsledku předchozího experimentu lze konstatovat, že tento způsob výběru dat není vhodný, proto je potřeba zvolit jiný způsob výběru dat. Místo řazení dat od nejmenšího k největšímu je zkoumána opačná varianta (sestupné řazení). Rozsah hodnot *F-skóre* pro vybírané koncepty je možné vidět v tabulce 5.7:

10% +	10%	20%	40% dat
vybráno náhodně	96,2 až 90,9	96,2 až 86,0	96,2 až 82,1

Tabulka 5.7: Rozsah hodnot F-skóre pro sestupně seřazená data.

Po spuštění experimentu pro sestupně seřazené hodnoty *F-skóre* bylo dosaženo výsledných hodnot konceptové přesnosti zobrazených v závislosti na množství vybíraných dat zobrazených na 5.7 :



Obrázek 5.7: Výběr dat dle F-skóre, data řazena sestupně.

Při sestupném řazení dat se výsledky obdržené touto strategií nezlepšují. Data ukazují podobné výsledky jako předchozí experiment. Pro $10 + 10\%$ vybíraných dat je dosaženo vyšší hodnoty *cAcc* než u *baseline*. V případě $10 + 20\%$ a $10 + 40\%$ dat dochází k viditelnému zhoršení výsledků. Výsledné hodnoty konceptové přesnosti se pohybují v rozsahu $76,59\%$ až $80,87\%$. Pro snazší orientaci jsou v tabulce 5.8 uvedeny výsledky obou experimentů i s jejich procentuálním nárůstem cAcc oproti baseline (I_m).

	Hodnota *cAcc* pro			I_m [%]
	$10 + 10\%$	$10 + 20\%$	$10 + 40\%$ dat	
Baseline	76,24	78,63	81,60	x
Výběr dle F-skóre (data vzestupně seřazená)	76,30	78,80	80,17	-0,49
Výběr dle F-skóre (data sestupně seřazená)	76,59	78,31	80,87	-0,25

Tabulka 5.8: Souhrnná tabulka pro výběr dat dle F-skóre.

Strategií výběru vstupních trénovacích dat dle hodnoty *F-skóre* není dosaženo průměrného procentuálního zvýšení konceptové přesnosti oproti *baseline*. Pro tento způsob výběru dat není ani pro jeden experiment potvrzen počáteční předpoklad (str. 26).

5.4 Výběr dat dle slovníku

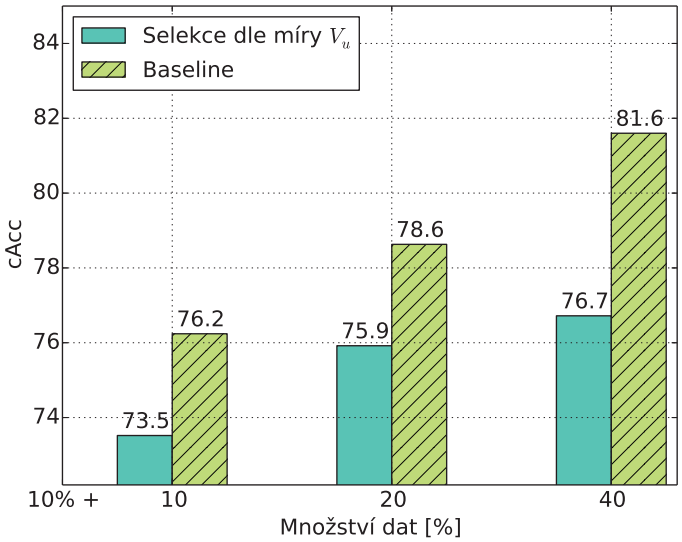
Poslední zkoumaná možnost výběru vstupních tréninkových dat je za pomoci slovníku $\mathcal{V}$ vytvořeného z 10% předepsaných dat. Pro zbylých 90% dat jsou jednotlivé promluvy ohodnoceny mírou V_u blíže popsanou v kapitole: *„Míra V_u“*. K ohodnocení je použit vytvořený slovník $\mathcal{V}$. Konkrétnější popis této strategie výběru dat naleznete v kapitole: *„Strategie výběru dat dle jejich slovníku“*.

V prvním experimentu byla ohodnocená data seřazena od největší hodnoty V_u k nejmenší. Tato hodnota vypovídá o tom, jaké množství slov v promluvě se již vyskytovalo ve slovníku $\mathcal{V}$. Pokud se všechna slova z promluvy vyskytují ve slovníku $\mathcal{V}$ hodnota *V-míry* je 1. Pokud naopak žádné ze slov není ve slovníku má hodnotu 0. Hodnoty výběrového parametru jsou pro výběr 10+10%, 10+20% a 10+40% dat jsou následující:

10% +	10%	20%	40% dat
vybráno náhodně	1,0 až 1,0	1,0 až 1,0	1,0 až 1,0

Tabulka 5.9: Rozsah hodnot V_u pro sestupně seřazená data.

Po spuštění experimentu pro sestupně seřazená data bylo dosaženo výsledků zobrazených v grafu 5.8:



Obrázek 5.8: Výběr dat dle V_u, data řazena sestupně.

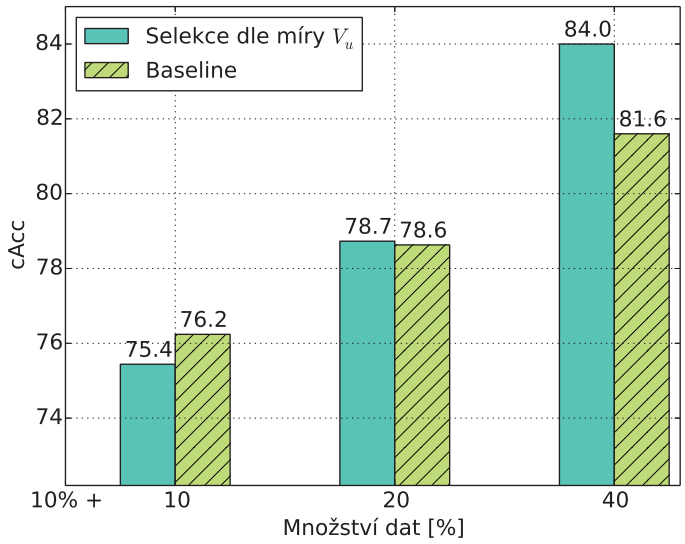
Z tabulky 5.9 je patrné, že jsou vybírány promluvy, jejichž všechna slova se vyskytují ve slovníku $\mathcal{V}$. Na základě získaných výsledků v grafu 5.8 jde o data, která nepřinášejí do procesu trénování žádné nové informace. V porovnání s původní *baseline*, jsou všechny hodnoty *cAcc* výrazně nižší a pohybují se v rozsahu $74,52\%$ až $76,72\%$.

Tento experimentu ukázal, že způsob výběru dat dle V_u má velký vliv na výslednou hodnotu konceptovou přesnost. Tento poznatek nás vede k dalšímu experimentování s tímto faktorem výběru. Dále je prozkoumána možnost řazení od nejmenší k největší hodnotě tohoto parametru. V tabulce 5.10 je viditelné, že pro 40% dat je hodnota parametru V_u rovna 1. Tento výsledek prozrazuje, že pro více jak 44% promluv z ohodnocených 90% dat jsou všechna slova obsažena ve slovníku $\mathcal{V}$.

10% +	10%	20%	40% dat
vybráno náhodně	0,0 až 0,5	0,0 až 0,666667	0,0 až 1,0

Tabulka 5.10: Rozsah hodnot V_u pro vzestupně seřazená data.

Při spuštění experimentu pro výběr vzestupně seřazených dat dle hodnoty V_u bylo dosaženo výsledků zobrazených v grafu na 5.8:



Obrázek 5.9: Výběr dat dle V_u, data řazena vzestupně.

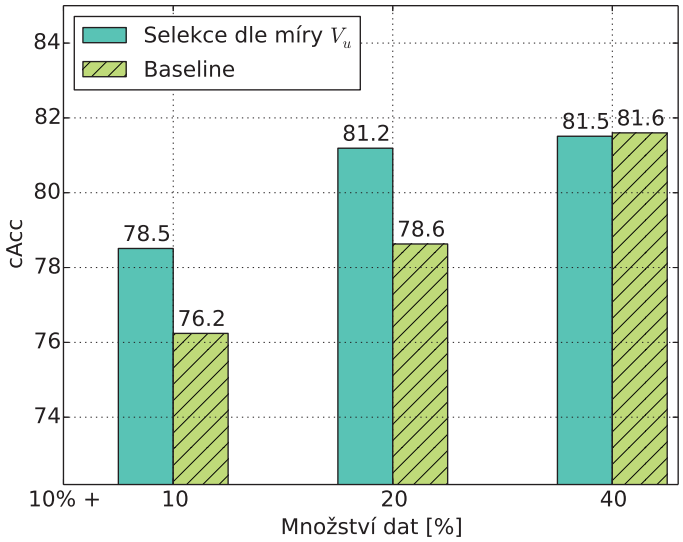
Z výsledků tohoto experimentu zobrazeného v grafu 5.9 je pozorovatelné, že pro $10 + 10\%$ vstupních dat je v porovnání s baseline dosaženo nižších hodnot konceptové přesnosti, tyto výsledky jsou pro promluvy s hodnotou *V-míry* v rozsahu 0 až $0,5$. Tento rozsah hodnot odpovídá promluvám, jejichž slova se ve slovníku $\mathcal{V}$ vyskytují z méně než 50%. Pro $10 + 20\%$ dat je hodnota konceptové přesnosti podobná hodnotě *baseline*, ale v porovnání s předchozí hodnotou *cAcc* je znatelně vyšší. Pro celý experiment se hodnota *cAcc* pohybuje v rozsahu $75,14\%$ až $84,00\%$.

Nejlepších výsledků je dosaženo výběrem dat obsahující promluvy s hodnotou V_u v rozsahu $0,666$ až $1,0$. U následujícího experimentu se proto soustředíme na oříznutí seřazených trénovacích dat s hodnotou V_u menší než $0,7$. Tomuto požadavku odpovídá prvních 25%. Nové hodnoty míry V_u je možné vidět v následující tabulce:

10% +	10%	20%	40% dat
vybráno náhodně	0,7058 až 0,5	0,7058 až 0,833	0,7058 až 1,0

Tabulka 5.11: Rozsah hodnot V_u pro oříznutá data (25%).

Po natrénování modelu HDM vzestupně seřazených a oříznutých dat bylo dosaženo výsledků zobrazených v grafu na 5.10:



Obrázek 5.10: Výběr dat dle V_u, data řazena vzestupně (ořez 25%).

Vyhodnocením experimentů jsou získány výsledky zobrazené v grafu 5.10. Pro prvních $10 + 20\%$ dat je dosaženo vyšší hodnoty konceptové přesnosti než v případě *baseline*. Pro $10 + 40\%$ dat dochází k vyčerpání trénovací množiny a výsledky porovnávaných strategií jsou velmi podobné. Výsledná konceptová přesnost se pohybuje v rozsahu $78,51\%$ až $81,51\%$. Pro názornější porovnání obou strategií jsou v tabulce 5.12 uvedeny výsledné hodnoty *cAcc* spolu s procentuálním nárůstem této míry oproti *baseline*.

	Hodnota *cAcc* pro			I_m [%]
	$10 + 10\%$	$10 + 20\%$	$10 + 40\%$ dat	
Baseline	76,24	78,63	81,60	x
Výběr dat dle V_u (data sestupně seřazená)	73,52	75,92	76,72	-4,32
Výběr dat dle V_u (data vzestupně seřazená)	75,44	78,73	84,00	0,66
Výběr dat dle V_u (data vzestupně seřazená, ořez 25%)	78,51	81,19	81,51	2,01

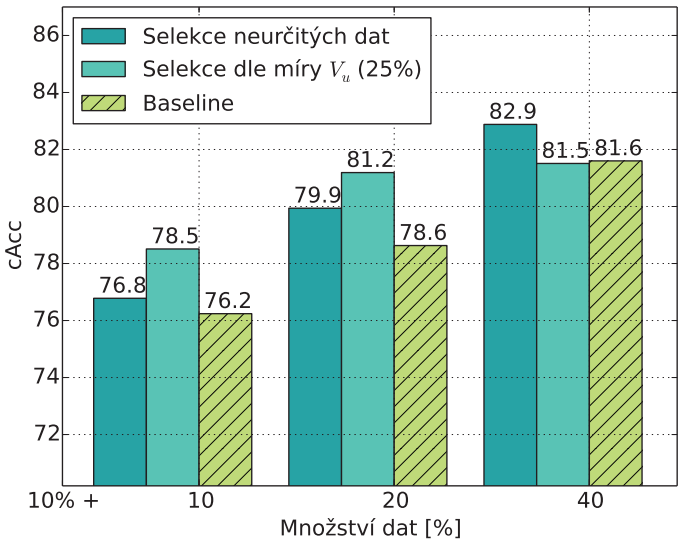
Tabulka 5.12: Souhrnná tabulka pro výběr dat dle V_u.

Experimenty pro výběr vstupních dat dle slovníku $\mathcal{V}$ vytvořeného z předepsaných 10% dat přinesla velmi dobré výsledky. Při oříznutí první čtvrtiny dat byly dokonce dosaženo průměrného procentuálního nárůstu 2,01% vyšší než u *baseline*. Experimentální ověření této strategie pro vzestupně seřazená data (bez a s ořezem 25%) potvrzují předpoklad definovaný v úvodu této kapitoly (str. 5).

5.5 Závěrečné porovnání strategií

Na základě provedených experimentů lze potvrdit počáteční předpoklad (str. 5). Vhodnějším výběrem vstupních dat pro trénování modelu porozumění řeči lze dosáhnout zvýšení míry *cAcc*, která slouží k vyhodnocení přesnosti modelu porozumění mluvené řeči.

Nejvyššího průměrného nárůstu konceptové přesnosti v porovnání s *baseline* bylo dosaženo použitím strategií 5.2 a 5.4. Pro strategii výběru dat dle jejich neurčitosti byl vyhodnocen jako nejvíce přínosný experiment vybírající z množiny vzestupně seřazených (neurčitých) dat. Řazení probíhalo na základě hodnoty posteriorní pravděpodobnosti výsledného sémantického stromu. Ve strategii výběru dat dle jejich slovníku byl jako nejlepší vyhodnocen experiment využívající vzestupně seřazená data dle míry V_u s oříznutím prvních 25% dat. Důsledkem oříznutí dat je pro množství $10 + 40\%$ dat hodnota konceptové přesnosti nižší. Tento pokles je způsoben vyčerpáním množiny vhodných dat pro trénování. Porovnání těchto dvou experimentů s baseline lze nalézt v grafu: 5.11.



Obrázek 5.11: Porovnání výsledků experimentů s nejvyšší hodnotou *cAcc*.

Kapitola 6

Závěr

V této diplomové práci byl zkoumán vliv výběru dat pro trénování modelu porozumění řeči na výslednou konceptovou přesnost modelu po jeho natrénování. Experimentální ověření navržených strategií, popsané v kapitole 5, byla provedena pro model HDM[5]. Trénovací data byla vybírána ze sad řečového korpusu TIA. Na základě těchto ověření je pro strategie dle *neurčitosti* a *slovníku* potvrzen počáteční předpoklad (str. 5). Při vhodném výběru trénovacích dat je v nejlepším experimentu (výběr ze vzestupně seřazených, oříznutých (25%) dat dle V_u) zvýšena průměrná přesnost porozumění mluvené řeči z referenčního slovního přepisu v porovnání s *baseline* o 2,02%.

Výsledky prováděných experimentů mohou být velmi užitečné v praxi. Vhodným výběrem dat lze dosáhnout zlepšením konceptové přesnosti. V případě velkého množství reálných dat je důležité umět správně rozhodnout jaká data pro trénování modelu vybrat. Správný výběr dat je důležitý pro zvýšení přesnosti porozumění mluvené řeči natrénovaného modelu. Tímto vhodným výběrem je možné dosáhnout snížení nákladů.

První navržená strategie (kapitola 4.4) odpovídá na otázku volby druhu vstupních trénovacích dat. Metodou byla ověřena *manuální*, *reálná* a kombinace *manuálních a reálných* dat. Výsledky experimentálních ověření (kapitola 5.1) ukazují, že vyšších hodnot konceptové přesnosti je dosaženo kombinací manuálních a reálných dat. Na základě těchto výsledků byla tato kombinace dat použita k experimentálnímu ověření strategií pro výběr trénovacích dat dle jejich *neurčitosti*, *špatné predikci* a *slovníku*.

Druhá navržená strategie (kapitola 4.5) zkoumá možnosti výběru vhodných dat pro anotaci na základě hodnoty jejich posteriorní pravděpodobnosti. Experimentální ověření

této strategie (kapitola 5.2) bylo provedeno pro seřazená data (vzestupně, sestupně). V případě vzestupného řazení byla navíc oříznuta o prvních 7.5% a 50% procent dat. Nejlepších výsledků bylo dosaženo experimenty pro vzestupné řazení dat (průměrný nárůst *cAcc* oproti *baseline* $1,31\%$) a vzestupné řazení dat s ořezem prvních 7.5% ($I_m = 0,81\%$).

Třetí navržená strategie (kapitola 4.5) se zabývá výběrem vhodných dat pro trénování modelu porozumění mluvené řeči dle hodnoty F-skóre. Výsledky experimentů (kapitola 5.3) provedené pro vzestupné a sestupné řazení dat ukazují, že při použití této strategie dochází v porovnání s *baseline* k poklesu konceptové přesnosti *cAcc*.

Čtvrtá navržená strategie (kapitola 4.6) představuje možnost výběru trénovacích dat za pomoci hodnoty výběrového faktoru V_u. Experimentálním ověření této strategie (kapitola 5.4) bylo provedeno pro seřazená data (vzestupně, sestupně). V případě vzestupného řazení navíc i s ořezem prvních 25% dat. Nejvyšší nárůst konceptové přesnosti byl zaznamenán u sestupného řazení dat s ořezem prvních 25% dat (průměrný nárůst cAcc oproti *baseline* o 2,01%) a při vzestupném řazení dat ($I_m = 0,66\%$).

Cílem této práce bylo maximalizovat přesnost porozumění mluvené řeči výběrem vstupních trénovacích dat. K výběru dat byly navrhnuty a ověřeny strategie výběru dat náhodně (*baseline*) a dle výběrového faktoru (*posteriorní pravděpodobnost*, *F-skóre* a V_u). Výsledky experimentálních ověření strategií (výběr dat dle *posteriorní pravděpodobnost* a V_u) kladně potvrdily zvýšení konceptové přesnosti v porovnání s *baseline*.

Literatura

[1] J. Psutka, L. Müller, J. Matoušek, and V. Radová. *Mluvíme s počítačem česky*. Academia, 2006. ISBN 80-200-1309-1.

[2] K. Loukotová. Uživatelská rozhraní dialogových systémů. Master's thesis, Univerzita Karlova v Praze, 2009. ISSN 1801-5948.

[3] A.M. Turing. Computing machinery and intelligence. In *Mind*, 1950, s. 433-460. ISSN: 00264423.

[4] M. Klocek. Základní vstupní a výstupní moduly pro voicexml interpreter jvoicexml. Technical report, Masarykova Univerzita, 2009.

[5] J. Švec. *Diskriminativní model pro porozumění mluvené řeči*. PhD thesis, Západočeská univerzita, Plzeň, 2013.

[6] F. Mairesse, M. Gašić, F. Jurčíček, et al. Spoken language understanding from unaligned data using discriminative classification models. *Acoustics, Speech and Signal Processing*, 2009, s. 4749-4752. IEEE International Conference on. IEEE, 2009. ISBN: 978-1-4244-2353-8.

[7] J. Švec. Sémantická analýza promluv systému NÁDRAŽÍ. Master's thesis, Západočeská univerzita, Plzeň, 2007.

[8] D. Jurafsky, D.H. Martin, et al. *Speech and language processing*. New York: Prentica Hall, 2000. ISBN: 978-0-13-187321-6.

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning 20.3*, 1995, s. 273-297.

[10] CH.C.J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery 2.2*, 1998, s. 121-167.

[11] D. Povey, M. Hannemann, et al. Generating exact lattices in the wfst framework. *IEEE International Conference on Acoustics Speech and Signal Processing, Kyoto, Japan*, 2002, s. 4213-4216.

[12] J. Vavruška. Metody pro úlohu vyhledávání slov v rozsáhlém archivu mluvené řeči. Master's thesis, Západočeská univerzita, Plzeň, 2012.

[13] A.B. Graf, A.J. Smola, and S. Borer. Classification in a normalized feature space using support vector machines. *Neural Networks*, 2003. IEEE Transactions on, 14(3), s. 597-605. ISSN: 0899-7667.

[14] C. Allauzen and M. et al. Riley. Openfst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, 2007, s. 11-23.

[15] Ting-Fan Wu, Chin jen Lin, and Weng R.C. Propability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 2004, s. 975-1005. ISSN: 0885-6125.